DOCUMENT RESUME

ED 477 483 IR 022 447

TITLE	Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide.
INSTITUTION	Institute of Education Sciences (ED), Washington, DC.
SPONS AGENCY	Council for Excellence in Government, Washington, DC.
PUB DATE	2003-12-00
NOTE	26p.; Prepared by the Institute for Education Sciences by the Coalition for Evidence-Based Policy, a project sponsored by the Council for Excellence in Government.
AVAILABLE FROM	Coalition for Evidence-Based Policy, 1301 K Street, NW, Suite 450 West, Washington, DC 20005. Tel: 202-728-0418; Fax: 202-728-0422; Web site: http://www.excelgov.org/evidence/. For full text: http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf/.
PUB TYPE	Guides - Non-Classroom (055) -- Reports - Descriptive (141)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Educational Improvement; *Educational Practices; *Educational Research; Elementary Secondary Education; *Instructional Effectiveness; *Research Design; Web Sites
IDENTIFIERS	*Evidence Based Practice

ABSTRACT

This guide provides assistance to K-12 educational practitioners in evaluating whether an educational intervention is backed by rigorous evidence of effectiveness, and in implementing evidence-based interventions in their schools or classrooms. The guide is organized in four parts: (1) a description of the randomized controlled trial and why it is a critical factor in establishing "strong" evidence of an intervention's effectiveness; (2) how to evaluate whether an intervention is backed by "strong" evidence of effectiveness; (3) how to evaluate whether an intervention is backed by "possible" evidence of effectiveness; and (4) important factors to consider when implementing an evidence-based intervention in schools or classrooms. Appendix A lists useful Web sites in finding evidence-based educational interventions. Appendix B provides a checklist to use in evaluating whether an intervention is backed by rigorous evidence. (Contains 26 references.) (MES)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 477 483

Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User Friendly Guide

U.S. Department of Education
Institute of Education Sciences
National Center for Education Evaluation
and Regional Assistance

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

IR022447

BEST COPY AVAILABLE

December 2003

Prepared for the:
Institute of Education Sciences
Grover J. Whitehurst, Director
by the

COALITION FOR EVIDENCE-BASED POLICY

A Project Sponsored by

THE COUNCIL FOR
EXCELLENCE
IN GOVERNMENT

COALITION BOARD OF ADVISORS

Robert Boruch
University of Pennsylvania

Jonathan Crane
Progressive Policy Institute

David Ellwood
Harvard University

Judith Gueron
Manpower Demonstration Research Corporation

Ron Haskins
Brookings Institution

Robert Hoyt
Jennison Associates

David Kessler
University of California, San Francisco

Jerry Lee
WIEB 101.1 FM Philadelphia

Diane Ravitch
New York University

Laurie Robinson
University of Pennsylvania

Isabel Sawhill
Brookings Institution

Martin Seligman
University of Pennsylvania

Robert Slavin
Johns Hopkins University

Robert Solow
Massachusetts Institute of Technology

Nicholas Zill
Westat, Inc.

EXECUTIVE DIRECTOR

Jon Baron
jbaron@excelgov.org

1301 K Street, NW
Suite 450 West
Washington, DC 20005
202-728-0418
FAX 202-728-0422
www.excelgov.org/evidence

BEST COPY AVAILABLE

Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User Friendly Guide

Purpose and Executive Summary iii

Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User-Friendly Guide 1

I. The randomized controlled trial: What it is, and why it is a critical factor in establishing "strong" evidence of an intervention's effectiveness. 1

II. How to evaluate whether an intervention is backed by "strong" evidence of effectiveness. 5

III. How to evaluate whether an intervention is backed by "possible" evidence of effectiveness. 11

IV. Important factors to consider when implementing an evidence-based intervention in your schools or classrooms. 13

Appendix A: Where to find evidence-based interventions 15

Appendix B: Checklist to use in evaluating whether an intervention is backed by rigorous evidence 16

References 18

BEST COPY AVAILABLE

Purpose and Executive Summary

This Guide seeks to provide educational practitioners with user-friendly tools to distinguish practices supported by rigorous evidence from those that are not.

The field of K-12 education contains a vast array of educational interventions – such as reading and math curricula, schoolwide reform programs, after-school programs, and new educational technologies – that claim to be able to improve educational outcomes and, in many cases, to be supported by evidence. This evidence often consists of poorly-designed and/or advocacy-driven studies. State and local education officials and educators must sort through a myriad of such claims to decide which interventions merit consideration for their schools and classrooms. Many of these practitioners have seen interventions, introduced with great fanfare as being able to produce dramatic gains, come and go over the years, yielding little in the way of positive and lasting change – a perception confirmed by the flat achievement results over the past 30 years in the National Assessment of Educational Progress long-term trend.

The federal *No Child Left Behind Act of 2001*, and many federal K-12 grant programs, call on educational practitioners to use "scientifically-based research" to guide their decisions about which interventions to implement. As discussed below, we believe this approach can produce major advances in the effectiveness of American education. Yet many practitioners have not been given the tools to distinguish interventions supported by scientifically-rigorous evidence from those which are not. This Guide is intended to serve as a user-friendly resource that the education practitioner can use to identify and implement evidence-based interventions, so as to improve educational and life outcomes for the children they serve.

If practitioners have the tools to identify evidence-based interventions, they may be able to spark major improvements in their schools and, collectively, in American education.

As illustrative examples of the potential impact of evidence-based interventions on educational outcomes, the following have been found to be effective in randomized controlled trials – research's "gold standard" for establishing what works:

- **One-on-one tutoring by qualified tutors for at-risk readers in grades 1-3** (the average tutored student reads more proficiently than approximately 75% of the untutored students in the control group).[1]
- **Life-Skills Training for junior high students** (low-cost, replicable program reduces smoking by 20% and serious levels of substance abuse by about 30% by the end of high school, compared to the control group).[2]
- **Reducing class size in grades K-3** (the average student in small classes scores higher on the Stanford Achievement Test in reading/math than about 60% of students in regular-sized classes).[3]
- **Instruction for early readers in phonemic awareness and phonics** (the average student in these interventions reads more proficiently than approximately 70% of students in the control group).[4]

In addition, preliminary evidence from randomized controlled trials suggests the effectiveness of:

- **High-quality, educational child care and preschool for low-income children** (by age 15, reduces special education placements and grade retentions by nearly 50% compared to controls; by age 21, more than doubles the proportion attending four-year college and reduces the percentage of teenage parents by 44%).[5]

BEST COPY AVAILABLE

Further research is needed to translate this finding into broadly-replicable programs shown effective in typical classroom or community settings.

The fields of medicine and welfare policy show that practice guided by rigorous evidence can produce remarkable advances.

Life and health in America has been profoundly improved over the past 50 years by the use of medical practices demonstrated effective in randomized controlled trials. These research-proven practices include: (i) vaccines for polio, measles, and hepatitis B; (ii) interventions for hypertension and high cholesterol, which have helped bring about a decrease in coronary heart disease and stroke by more than 50 percent over the past half-century; and (iii) cancer treatments that have dramatically improved survival rates from leukemia, Hodgkin's disease, and many other types of cancer.

Similarly, welfare policy, which since the mid-1990s has been remarkably successful in moving people from welfare into the workforce, has been guided to a large extent by scientifically-valid knowledge about "what works" generated in randomized controlled trials.[6]

Our hope is that this Guide, by enabling educational practitioners to draw effectively on rigorous evidence, can help spark similar evidence-driven progress in the field of education.

The diagram on the next page summarizes the process we recommend for evaluating whether an educational intervention is supported by rigorous evidence.

In addition, appendix B contains a checklist to use in this process.

BEST COPY AVAILABLE

How to evaluate whether an educational intervention is supported by rigorous evidence: An overview

Step 1. **Is the intervention backed by "strong" evidence of effectiveness?**

Quality of studies needed to establish "strong" evidence:

- Randomized controlled trials (defined on page 1) that are well-designed and implemented (see pages 5-9).

+

Quantity of evidence needed:

Trials showing effectiveness in —

- Two or more typical school settings,
- Including a setting similar to that of your schools/ classrooms.

(see page 10)

=

"Strong" Evidence

Step 2. **If the intervention is not backed by "strong" evidence, is it backed by "possible" evidence of effectiveness?**

Types of studies that can comprise "possible" evidence:

- Randomized controlled trials whose quality/quantity are good but fall short of "strong" evidence (see page 11); and/or
- Comparison-group studies (defined on page 3) in which the intervention and comparison groups are *very closely matched* in academic achievement, demographics, and other characteristics (see pages 11-12).

Types of studies that do not comprise "possible" evidence:

- Pre-post studies (defined on page 2).
- Comparison-group studies in which the intervention and comparison groups are not closely matched (see pages 12-13).
- "Meta-analyses" that include the results of such lower-quality studies (see page 13).

Step 3. **If the answers to both questions above are "no," one may conclude that the intervention is not supported by meaningful evidence.**

BEST COPY AVAILABLE

Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User-Friendly Guide

This Guide seeks to provide assistance to educational practitioners in evaluating whether an educational intervention is backed by rigorous evidence of effectiveness, and in implementing evidence-based interventions in their schools or classrooms. By *intervention*, we mean an educational practice, strategy, curriculum, or program. The Guide is organized in four parts:

I. A description of the randomized controlled trial, and why it is a critical factor in establishing "strong" evidence of an intervention's effectiveness;

II. How to evaluate whether an intervention is backed by "strong" evidence of effectiveness;

III. How to evaluate whether an intervention is backed by "possible" evidence of effectiveness; and

IV. Important factors to consider when implementing an evidence-based intervention in your schools or classrooms.

I. The randomized controlled trial: What it is, and why it is a critical factor in establishing "strong" evidence of an intervention's effectiveness.

Well-designed and implemented randomized controlled trials are considered the "gold standard" for evaluating an intervention's effectiveness, in fields such as medicine, welfare and employment policy, and psychology.[7] This section discusses what a randomized controlled trial is, and outlines evidence indicating that such trials should play a similar role in education.

A. Definition: Randomized controlled trials are studies that randomly assign individuals to an intervention group or to a control group, in order to measure the effects of the intervention.

For example, suppose you want to test, in a randomized controlled trial, whether a new math curriculum for third-graders is more effective than your school's existing math curriculum for third-graders. You would randomly assign a large number of third-grade students to either an intervention group, which uses the new curriculum, or to a control group, which uses the existing curriculum. You would then measure the math achievement of both groups over time. The difference in math achievement between the two groups would represent the effect of the new curriculum compared to the existing curriculum.

In a variation on this basic concept, sometimes individuals are randomly assigned to two or more intervention groups as well as to a control group, in order to measure the effects of different interventions in one trial. Also, in some trials, entire classrooms, schools, or school districts – rather than individual students – are randomly assigned to intervention and control groups.

BEST COPY AVAILABLE

B. <u>The unique advantage of random assignment</u>: It enables you to evaluate whether the intervention itself, as opposed to other factors, causes the observed outcomes.

Specifically, the process of randomly assigning a large number of individuals to either an intervention group or a control group ensures, to a high degree of confidence, that there are no systematic differences between the groups in any characteristics (observed and unobserved) except one – namely, the intervention group participates in the intervention, and the control group does not. Therefore – assuming the trial is properly carried out (per the guidelines below) – the resulting difference in outcomes between the intervention and control groups can confidently be attributed to the intervention and not to other factors.

C. There is persuasive evidence that the randomized controlled trial, when properly designed and implemented, is superior to other study designs in measuring an intervention's true effect.

1. "Pre-post" study designs often produce erroneous results.

<u>Definition</u>: A "pre-post" study examines whether participants in an intervention improve or regress during the course of the intervention, and then attributes any such improvement or regression to the intervention.

The problem with this type of study is that, without reference to a control group, it cannot answer whether the participants' improvement or decline would have occurred anyway, even without the intervention. This often leads to erroneous conclusions about the effectiveness of the intervention.

<u>Example</u>: A randomized controlled trial of Even Start – a federal program designed to improve the literacy of disadvantaged families – found that the program had no effect on improving the school readiness of participating children at the 18th-month follow-up. Specifically, there were no significant differences between young children in the program and those in the control group on measures of school readiness including the Picture Peabody Vocabulary Test (PPVT) and PreSchool Inventory.[8]

If a pre-post design rather than a randomized design had been used in this study, the study would have concluded erroneously that the program was effective in increasing school readiness. This is because *both* the children in the program and those in the control group showed improvement in school readiness during the course of the program (e.g., both groups of children improved substantially in their national percentile ranking on the PPVT). A pre-post study would have attributed the participants' improvement to the program whereas in fact it was the result of other factors, as evidenced by the equal improvement for children in the control group.

<u>Example</u>: A randomized controlled trial of the Summer Training and Education Program – a Labor Department pilot program that provided summer remediation and work experience for disadvantaged teenagers – found that program's short-term impact on participants' reading ability was positive. Specifically, while the reading ability of the control group members eroded by a full grade-level during the first summer of the program, the reading ability of participants in the program eroded by only a half grade-level.[9]

If a pre-post design rather than a randomized design had been used in this study, the study would have concluded erroneously that the program was harmful. That is, the study would have found a decline in participants' reading ability and attributed it to the program. In fact, however, the participants' decline in reading ability was the result of other factors – such as the natural erosion of reading ability during the summer vacation months – as evidenced by the even greater decline for members of the control group.

BEST COPY AVAILABLE

2. **The most common "comparison group" study designs (also known as "quasi-experimental" designs) also lead to erroneous conclusions in many cases.**

   a. **Definition: A "comparison group" study compares outcomes for intervention participants with outcomes for a comparison group chosen through methods other than randomization.**

      The following example illustrates the basic concept of this design. Suppose you want to use a comparison-group study to test whether a new mathematics curriculum is effective. You would compare the math performance of students who participate in the new curriculum ("intervention group") with the performance of a "comparison group" of students, chosen through methods other than randomization, who do not participate in the curriculum. The comparison group might be students in neighboring classrooms or schools that don't use the curriculum, or students in the same grade and socioeconomic status selected from state or national survey data. The difference in math performance between the intervention and comparison groups following the intervention would represent the estimated effect of the curriculum.

      Some comparison-group studies use statistical techniques to create a comparison group that is matched with the intervention group in socioeconomic and other characteristics, or to otherwise adjust for differences between the two groups that might lead to inaccurate estimates of the intervention's effect. The goal of such statistical techniques is to simulate a randomized controlled trial.

   b. **There is persuasive evidence that the most common comparison-group designs produce erroneous conclusions in a sizeable number of cases.**

      A number of careful investigations have been carried out – in the areas of school dropout prevention,[10] K-3 class-size reduction,[11] and welfare and employment policy[12] – to examine whether and under what circumstances comparison-group designs can replicate the results of randomized controlled trials.[13] These investigations first compare participants in a particular intervention with a control group, selected through randomization, in order to estimate the intervention's impact in a randomized controlled trial. Then the same intervention participants are compared with a comparison group selected through methods other than randomization, in order to estimate the intervention's impact in a comparison-group design. Any systematic difference between the two estimates represents the inaccuracy produced by the comparison-group design.

      These investigations have shown that most comparison-group designs in education and other areas produce inaccurate estimates of an intervention's effect. This is because of unobservable differences between the members of the two groups that differentially affect their outcomes. For example, if intervention participants self-select themselves into the intervention group, they may be more motivated to succeed than their control-group counterparts. Their motivation – rather than the intervention – may then lead to their superior outcomes. In a sizeable number of cases, the inaccuracy produced by the comparison-group designs is large enough to result in erroneous overall conclusions about whether the intervention is effective, ineffective, or harmful.

Example from medicine. Over the past 30 years, more than two dozen comparison-group studies have found hormone replacement therapy for postmenopausal women to be effective in reducing the women's risk of coronary heart disease, by about 35-50 percent. But when hormone therapy was finally evaluated in two large-scale randomized controlled trials – medicine's "gold standard" – it was actually found to do the opposite: it *increased* the risk of heart disease, as well as stroke and breast cancer.[14]

Medicine contains many other important examples of interventions whose effect as measured in comparison-group studies was subsequently contradicted by well-designed randomized controlled trials. If randomized controlled trials in these cases had never been carried out and the comparison-group results had been relied on instead, the result would have been needless death or serious illness for millions of people. This is why the Food and Drug Administration and National Institutes of Health generally use the randomized controlled trial as the final arbiter of which medical interventions are effective and which are not.

3. ***Well-matched* comparison-group studies can be valuable in generating hypotheses about "what works," but their results need to be confirmed in randomized controlled trials.**

   The investigations, discussed above, that compare comparison-group designs with randomized controlled trials generally support the value of comparison-group designs in which the comparison group is *very closely matched* with the intervention group in prior test scores, demographics, time period in which they are studied, and methods used to collect outcome data. In most cases, such well-matched comparison-group designs seem to yield correct overall conclusions in most cases about whether an intervention is effective, ineffective, or harmful. However, their estimates of the size of the intervention's impact are still often inaccurate. As an illustrative example, a well-matched comparison-group study might find that a program to reduce class size raises test scores by 40 percentile points – or, alternatively, by 5 percentile points – when its true effect is 20 percentile points. Such inaccuracies are large enough to lead to incorrect overall judgments about the policy or practical significance of the intervention in a nontrivial number of cases.

   As discussed in section III of this Guide, we believe that such well-matched studies can play a valuable role in education, as they have in medicine and other fields, in establishing "possible" evidence an intervention's effectiveness, and thereby generating hypotheses that merit confirmation in randomized controlled trials. But the evidence cautions strongly against using even the most well-matched comparison-group studies as a final arbiter of what is effective and what is not, or as a reliable guide to the strength of the effect.

D. Thus, we believe there are compelling reasons why randomized controlled trials are a critical factor in establishing "strong" evidence of an intervention's effectiveness.

BEST COPY AVAILABLE

II. How to evaluate whether an intervention is backed by "strong" evidence of effectiveness.

This section discusses how to evaluate whether an intervention is backed by "strong" evidence that it will improve educational outcomes in your schools or classrooms. Specifically, it discusses both the quality and quantity of studies needed to establish such evidence.

A. Quality of evidence needed to establish "strong" evidence of effectiveness: Randomized controlled trials that are well-designed and implemented.

As discussed in section I, randomized controlled trials are a critical factor in establishing "strong" evidence of an intervention's effectiveness. Of course, such trials must also be well-designed and implemented in order to constitute strong evidence. Below is an outline of key items to look for when reviewing a randomized controlled trial of an educational intervention, to see whether the trial was well-designed and implemented. It is meant as a discussion of general principles, rather than as an exhaustive list of the features of such trials.

Key items to look for in the study's description of the intervention and the random assignment process

1. The study should clearly describe (i) the intervention, including who administered it, who received it, and what it cost; (ii) how the intervention differed from what the control group received; and (iii) the logic of how the intervention is supposed to affect outcomes.

Example. A randomized controlled trial of a one-on-one tutoring program for beginning readers should discuss such items as:

- who conducted the tutoring (e.g., certified teachers, paraprofessionals, or undergraduate volunteers);
- what training they received in how to tutor;
- what curriculum they used to tutor, and other key features of the tutoring sessions (e.g., daily 20-minute sessions over a period of six-months);
- the age, reading achievement levels, and other relevant characteristics of the tutored students and controls;
- the cost of the tutoring intervention per student;
- the reading instruction received by the students in the control group (e.g., the school's pre-existing reading program); and
- the logic by which tutoring is supposed to improve reading outcomes.

2. Be alert to any indication that the random assignment process may have been compromised.

For example, did any individuals randomly assigned to the control group subsequently cross over to the intervention group? Or did individuals unhappy with their prospective assignment to either the intervention or control group have an opportunity to delay their entry into the study until another

BEST COPY AVAILABLE

opportunity arose for assignment to their preferred group? Such self-selection of individuals into their preferred groups undermines the random assignment process, and may well lead to inaccurate estimates of the intervention's effects.

Ideally, a study should describe the method of random assignment it used (e.g., coin toss or lottery), and what steps were taken to prevent undermining (e.g., asking an objective third party to administer the random assignment process). In reality, few studies – even well-designed trials – do this. But we recommend that you be alert to any indication that the random assignment process was compromised.

3. The study should provide data showing that there were no systematic differences between the intervention and control groups before the intervention.

As discussed above, the random assignment process ensures, to a high degree of confidence, that there are no systematic differences between the characteristics of the intervention and control groups prior to the intervention. However, in rare cases – particularly in smaller trials – random assignment might by chance produce intervention and control groups that differ systematically in various characteristics (e.g., academic achievement levels, socioeconomic status, ethnic mix). Such differences could lead to inaccurate results. Thus, the study should provide data showing that, before the intervention, the intervention and control groups did not differ systematically in the vast majority of measured characteristics (allowing that, by chance, there might have been some minor differences).

Key items to look for in the study's collection of outcome data

4. The study should use outcome measures that are "valid" – – i.e., that accurately measure the true outcomes that the intervention is designed to affect. Specifically:

- **To test academic achievement outcomes (e.g., reading/math skills), a study should use tests whose ability to accurately measure true skill levels is well-established** (for example, the Woodcock-Johnson Psychoeducational Battery, the Stanford Achievement Test, etc.).

- **Wherever possible, a study should use objective, "real-world" measures of the outcomes that the intervention is designed to affect** (e.g., for a delinquency prevention program, the students' official suspensions from school).

- **If outcomes are measured through interviews or observation, the interviewers/observers preferably should be kept unaware of who is in the intervention and control groups.**

  Such "blinding" of the interviewers/observers, where possible, helps protect against the possibility that any bias they may have (e.g., as proponents of the intervention) could influence their outcome measurements. Blinding would be appropriate, for example, in a study of a violence prevention program for elementary school students, where an outcome measure is the incidence of hitting on the playground as detected by an adult observer.

- **When study participants are asked to "self-report" outcomes, their reports should, if possible, be corroborated by independent and/or objective measures.**

  For instance, when participants in a substance-abuse or violence prevention program are asked to self-report their drug or tobacco use or criminal behavior, they tend to under-report such undesir-

BEST COPY AVAILABLE

able behaviors. In some cases, this may lead to inaccurate study results, depending on whether the intervention and control groups under-report by different amounts.

Thus, studies that use such self-reported outcomes should, if possible, corroborate them with other measures (e.g., saliva thiocyanate tests for smoking, official arrest data, third-party observations).

5. The percent of study participants that the study has lost track of when collecting outcome data should be small, and should not differ between the intervention and control groups.

A general guideline is that the study should lose track of fewer than 25 percent of the individuals originally randomized – the fewer lost, the better. This is sometimes referred to as the requirement for "low attrition." (Studies that choose to follow only a representative subsample of the randomized individuals should lose track of less than 25 percent of the subsample.)

Furthermore, the percentage of subjects lost track of should be approximately the same for the intervention and the control groups. This is because differential losses between the two groups can create systematic differences between the two groups, and thereby lead to inaccurate estimates of the intervention's effect. This is sometimes referred to as the requirement for "no differential attrition."

6. The study should collect and report outcome data even for those members of the intervention group who don't participate in or complete the intervention.

This is sometimes referred to as the study's use of an "intention-to-treat" approach, the importance of which is best illustrated with an example.

Example. Consider a randomized controlled trial of a school voucher program, in which students from disadvantaged backgrounds are randomly assigned to an intervention group – whose members are offered vouchers to attend private school – or to a control group that does not receive voucher offers. It's likely that some of the students in the intervention group will not accept their voucher offers and will choose instead to remain in their existing schools. Suppose that, as may well be the case, these students as a group are less motivated to succeed than their counterparts who accept the offer. If the trial then drops the students not accepting the offer from the intervention group, leaving the more motivated students, it would be create a systematic difference between the intervention and control groups – namely, motivation level. Thus the study may well over-estimate the voucher program's effect on educational success, erroneously attributing a superior outcome for the intervention group to the vouchers when in fact it was due to the difference in motivation.

Therefore, the study should collect outcome data for all of the individuals randomly assigned to the intervention group, *whether they participated in the intervention or not*, and should use all such data in estimating the intervention's effect. The study should also report on how many of the individuals assigned to the intervention group actually participated in the intervention.

7. The study should preferably obtain data on long-term outcomes of the intervention, so that you can judge whether the intervention's effects were sustained over time.

This is important because the effect of many interventions diminishes substantially within 2-3 years after the intervention ends. This has been demonstrated in randomized controlled trials in diverse areas such as early reading, school-based substance-abuse prevention, prevention of childhood

BEST COPY AVAILABLE

depression, and welfare-to-work and employment. In most cases, it is the longer-term effect, rather than the immediate effect, that is of greatest practical and policy significance.

Key items to look for in the study's reporting of results

8. If the study claims that the intervention improves one or more outcomes, it should report (i) the size of the effect, and (ii) statistical tests showing the effect is unlikely to be due to chance.

Specifically, the study should report the size of the difference in outcomes between the intervention and control groups. It should also report the results of tests showing the difference is "statistically significant" at conventional levels -- generally the .05 level. Such a finding means that there is only a 1 in 20 probability that the difference could have occurred by chance if the intervention's true effect is zero.

a. In order to obtain such a finding of statistically significant effects, a study usually needs to have a relatively large sample size.

A rough rule of thumb is that a sample size of at least 300 students (150 in the intervention group and 150 in the control group) is needed to obtain a finding of statistical significance for an intervention that is modestly effective. If schools or classrooms, rather than individual students, are randomized, a minimum sample size of 50 to 60 schools or classrooms (25-30 in the intervention group and 25-30 in the control group) is needed to obtain such a finding. (This rule of thumb assumes that the researchers choose a sample of individuals or schools/classrooms that do not differ widely in initial achievement levels.)[15] If an intervention is highly effective, smaller sample sizes than this may be able to generate a finding of statistical significance.

If the study seeks to examine the intervention's effect on particular subgroups within the overall sample (e.g., Hispanic students), larger sample sizes than those above may be needed to generate a finding of statistical significance for the subgroups.

In general, larger sample sizes are better than smaller sample sizes, because they provide greater confidence that any difference in outcomes between the intervention and control groups is due to the intervention rather than chance.

b. If the study randomizes groups (e.g., schools) rather than individuals, the sample size that the study uses in tests for statistical significance should be the number of groups rather than the number of individuals in those groups.

Occasionally, a study will erroneously use the number of individuals as its sample size, and thus generate false findings of statistical significance.

Example. If a study randomly assigns two schools to an intervention group and two schools to a control group, the sample size that the study should use in tests for statistical significance is just four, regardless of how many hundreds of students are in the schools. (And it is very unlikely that such a small study could obtain a finding of statistical significance.)

BEST COPY AVAILABLE

c. **The study should preferably report the size of the intervention's effects in easily understandable, real-world terms** (e.g., an improvement in reading skill by two grade levels, a 20 percent reduction in weekly use of illicit drugs, a 20 percent increase in high school graduation rates).

It is important for a study to report the size of the intervention's effects in this way, in addition to whether the effects are statistically significant, so that you (the reader) can judge their educational importance. For example, it is possible that a study with a large sample size could show effects that are statistically significant but so small that they have little practical or policy significance (e.g., a 2 point increase in SAT scores). Unfortunately, some studies report only whether the intervention's effects are statistically significant, and not their magnitude.

Some studies describe the size of the intervention's effects in "standardized effect sizes."[16] A full discussion of this concept is beyond the scope of this Guide. We merely comment that standardized effect sizes may not accurately convey the educational importance of an intervention, and, when used, should preferably be translated into understandable, real-world terms like those above.

9. A study's claim that the intervention's effect on a subgroup (e.g., Hispanic students) is different than its effect on the overall population in the study should be treated with caution.

Specifically, we recommend that you look for corroborating evidence of such subgroup effects in other studies before accepting them as valid.

This is because a study will sometimes show different effects for different subgroups just by chance, particularly when the researchers examine a large number of subgroups and/or the subgroups contain a small number of individuals. For example, even if an intervention's true effect is the same on all subgroups, we would expect a study's analysis of 20 subgroups to "demonstrate" a different effect on one of those subgroups just by chance (at conventional levels of statistical significance). Thus, studies that engage in a post-hoc search for different subgroup effects (as some do) will sometimes turn up spurious effects rather than legitimate ones.

Example. In a large randomized controlled trial of aspirin for the emergency treatment of heart attacks, aspirin was found to be highly effective, resulting in a 23 percent reduction in vascular deaths at the one-month follow-up. To illustrate the unreliability of subgroup analyses, these overall results were subdivided by the patients' astrological birth signs into 12 subgroups. Aspirin's effects were similar in most subgroups to those for the whole population. However, for two of the subgroups, Libra and Gemini, aspirin appeared to have no effect in reducing mortality. Clearly it would be wrong to conclude from this analysis that heart attack patients born under the astrological signs of Libra and Gemini do not benefit from aspirin. [17]

10. The study should report the intervention's effects on all the outcomes that the study measured, not just those for which there is a positive effect.

This is because if a study measures a large number of outcomes, it may, by chance alone, find positive (and statistically-significant) effects on one or a few of those outcomes. Thus, the study should report the intervention's effects on all measured outcomes so that you can judge whether the positive effects are the exception or the pattern.

BEST COPY AVAILABLE

B. Quantity of evidence needed to establish "strong" evidence of effectiveness.

1. For reasons set out below, we believe "strong" evidence of effectiveness requires:

(i) that the intervention be demonstrated effective, through well-designed randomized controlled trials, in more than one site of implementation, and

(ii) that these sites be typical school or community settings, such as public school classrooms taught by regular teachers. Typical settings would not include, for example, specialized classrooms set up and taught by researchers for purposes of the study.

Such a demonstration of effectiveness may require more than one randomized controlled trial of the intervention, or one large trial with more than one implementation site.

2. In addition, the trials should demonstrate the intervention's effectiveness in school settings similar to yours, before you can be confident it will work in your schools and classrooms.

For example, if you are considering implementing an intervention in a large inner-city public school serving primarily minority students, you should look for randomized controlled trials demonstrating the intervention's effectiveness in similar settings. Randomized controlled trials demonstrating its effectiveness in a white, suburban population do not constitute strong evidence that it will work in *your* school.

3. Main reasons why a demonstration of effectiveness in more than one site is needed:

- **A single finding of effectiveness can sometimes occur by chance alone.** For example, even if all educational interventions tested in randomized controlled trials were ineffective, we would expect 1 in 20 of those trials to "demonstrate" effectiveness by chance alone at conventional levels of statistical significance. Requiring that an intervention be shown effective in two trials (or in two sites of one large trial) reduces the likelihood of such a false-positive result to 1 in 400.

- **The results of a trial in any one site may be dependent on site-specific factors and thus may not be generalizable to other sites.** It is possible, for instance, that an intervention may be highly effective in a school with an unusually talented individual managing the details of implementation, but would not be effective in another school with other individuals managing the detailed implementation.

Example. Two multi-site randomized controlled trials of the Quantum Opportunity Program – a community-based program for disadvantaged high school students providing academic assistance, college and career planning, community service and work experiences, and other services – have found that the program's effects vary greatly among the various program sites. A few sites – including the original program site (Philadelphia) – produced sizeable effects on participants' academic and/or career outcomes, whereas many sites had little or no effect on the same outcomes.[18] Thus, the program's effects appear to be highly dependent on site-specific factors, and it is not clear that its success can be widely replicated.

4. Pharmaceutical medicine provides an important precedent for the concept that "strong" evidence requires a showing of effectiveness in more than one instance.

Specifically, the Food and Drug Administration (FDA) usually requires that a new pharmaceutical drug or medical device be shown effective in more than one randomized controlled trial before the FDA will grant it a license to be marketed. The FDA's reasons for this policy are similar to those discussed above.[19]

BEST COPY AVAILABLE

III. How to evaluate whether an intervention is backed by "possible" evidence of effectiveness.

Because well-designed and implemented randomized controlled trials are not very common in education, the evidence supporting an intervention frequently falls short of the above criteria for "strong" evidence of effectiveness in one or more respects. For example, the supporting evidence may consist of:

- Only nonrandomized studies;
- Only one well-designed randomized controlled trial showing the intervention's effectiveness at a single site;
- Randomized controlled trials whose design and implementation contain one or more flaws noted above (e.g., high attrition);
- Randomized controlled trials showing the intervention's effectiveness as implemented by researchers in a laboratory-like setting, rather.than in a typical school or community setting; or
- Randomized controlled trials showing the intervention's effectiveness for students with different academic skills and socioeconomic backgrounds than the students in your schools or classrooms.

Whether an intervention not supported by "strong" evidence is nevertheless supported by "possible" evidence of effectiveness (as opposed to *no* meaningful evidence of effectiveness) is a judgment call that depends, for example, on the extent of the flaws in the randomized controlled trials of the intervention and the quality of any nonrandomized studies that have been done. While this Guide cannot foresee and provide advice on all possible scenarios of evidence, it offers in this section a few factors to consider in evaluating whether an intervention not supported by "strong" evidence is nevertheless supported by "possible" evidence.

A. Circumstances in which a comparison-group study can constitute "possible" evidence of effectiveness:

1. The study's intervention and comparison groups should be very closely matched in academic achievement levels, demographics, and other characteristics prior to the intervention.

The investigations, discussed in section I, that compare comparison-group designs with randomized controlled trials generally support the value of comparison-group designs in which the comparison group is *very closely matched* with the intervention group. In the context of education studies, the two groups should be matched closely in characteristics including:

- Prior test scores and other measures of academic achievement (preferably, the same measures that the study will use to evaluate outcomes for the two groups);
- Demographic characteristics, such as age, sex, ethnicity, poverty level, parents' educational attainment, and single or two-parent family background;
- Time period in which the two groups are studied (e.g., the two groups are children entering kindergarten in the same year as opposed to sequential years); and
- Methods used to collect outcome data (e.g., the same test of reading skills administered in the same way to both groups).

These investigations have also found that when the intervention and comparison groups differ in such characteristics, the study is unlikely to generate accurate results even when statistical techniques are then used to adjust for these differences in estimating the intervention's effects.

2. **The comparison group should not be comprised of individuals who had the option to participate in the intervention but declined.**

   This is because individuals choosing not to participate in an intervention may differ systematically in their level of motivation and other important characteristics from the individuals who do choose to participate. The difference in motivation (or other characteristics) may itself lead to different outcomes for the two groups, and thus contaminate the study's estimates of the intervention's effects.

   Therefore, the comparison group should be comprised of individuals who did not have the option to participate in the intervention, rather than individuals who had the option but declined.

3. **The study should preferably choose the intervention/comparison groups and outcome measures "prospectively" – that is, *before* the intervention is administered.**

   This is because if the groups and outcomes measures are chosen by the researchers *after* the intervention is administered ("retrospectively"), the researchers may consciously or unconsciously select groups and outcome measures so as to generate their desired results. Furthermore, it is often difficult or impossible for the reader of the study to determine whether the researchers did so.

   Prospective comparison-group studies are, like randomized controlled trials, much less susceptible to this problem. In the words of the director of drug evaluation for the Food and Drug Administration, "The great thing about a [randomized controlled trial or prospective comparison-group study] is that, within limits, you don't have to believe anybody or trust anybody. The planning for [the study] is prospective; they've written the protocol before they've done the study, and any deviation that you introduce later is completely visible." By contrast, in a retrospective study, "you always wonder how many ways they cut the data. It's very hard to be reassured, because there are no rules for doing it."[20]

4. **The study should meet the guidelines set out in section II for a well-designed randomized controlled trial (other than guideline 2 concerning the random-assignment process).**

   That is, the study should use valid outcome measures, have low attrition, report tests for statistical significance, and so on.

B. **Studies that do *not* meet the threshold for "possible" evidence of effectiveness:**

1. **Pre-post studies, which often produce erroneous results, as discussed in section I.**

2. **Comparison-group studies in which the intervention and comparison groups are not well-matched.**

   As discussed in section I, such studies also produce erroneous results in many cases, even when statistical techniques are used to adjust for differences between the two groups.

Example. As reported in *Education Week*, several comparison-group studies have been carried out to evaluate the effects of "high-stakes testing" – i.e., state-level policies in which student test scores are used to determine various consequences, such as whether the students graduate or are promoted to the next grade, whether their teachers are awarded bonuses, or whether their school is taken over by the state. These studies compare changes in test scores and dropout rates for students in states with high-stakes testing (the intervention group) to those for students in other states (the comparison groups). Because students in different states differ in many characteristics, such as demographics and initial levels of academic achievement, it is unlikely that these studies provide accurate measures of the effects of high-stakes testing. It is not surprising that these studies reach differing conclusions about the effects of such testing.[21]

3. "Meta-analyses" that combine the results of individual studies that do not themselves meet the threshold for "possible" evidence.

Meta-analysis is a quantitative technique for combining the results of individual studies, a full discussion of which is beyond the scope of this Guide. We merely note that when meta-analysis is used to combine studies that themselves may generate erroneous results – such as randomized controlled trials with significant flaws, poorly-matched comparison group studies, and pre-post studies – it will often produce erroneous results as well.

Example. A meta-analysis combining the results of many nonrandomized studies of hormone replacement therapy found that such therapy significantly lowered the risk of coronary heart disease.[22] But, as noted earlier, when hormone therapy was subsequently evaluated in two large-scale randomized controlled trials, it was actually found to do the opposite – namely, it increased the risk of coronary disease. The meta-analysis merely reflected the inaccurate results of the individual studies, producing more precise, but still erroneous, estimates of the therapy's effect.

IV. Important factors to consider when implementing an evidence-based intervention in *your* schools or classrooms.

A. Whether an evidence-based intervention will have a positive effect in your schools or classrooms may depend critically on your adhering closely to the details of its implementation.

The importance of adhering to the details of an evidence-based intervention when implementing it in your schools or classrooms is often not fully appreciated. Details of implementation can sometimes make a major difference in the intervention's effects, as the following examples illustrate.

BEST COPY AVAILABLE

Example. The Tennessee Class-Size Experiment – a large, multi-site randomized controlled trial involving 12,000 students – showed that a state program that significantly reduced class size for public school students in grades K-3 had positive effects on educational outcomes. For example, the average student in the small classes scored higher on the Stanford Achievement Test in reading and math than about 60 percent of the students in the regular-sized classes, and this effect diminished only slightly at the fifth-grade follow-up.[23]

Based largely on these results, in 1996 the state of California launched a much larger, state-wide class-size reduction effort for students in grades K-3. But to implement this effort, California schools hired 25,000 new K-3 teachers, many with low qualifications. Thus the proportion of fully-credentialed K-3 teachers fell in most California schools, with the largest drop (16 percent) occurring in the schools serving the lowest-income students. By contrast, all the teachers in the Tennessee study were fully qualified. This difference in implementation may account for the fact that, according to preliminary comparison-group data, class-size reduction in California may not be having as large an impact as in Tennessee.[24]

Example. Three well-designed randomized controlled trials have established the effectiveness of the Nurse-Family Partnership – a nurse visitation program provided to low-income, mostly single women during pregnancy and their children's infancy. One of these studies included a 15-year follow-up, which found that the program reduced the children's arrests, convictions, number of sexual partners, and alcohol use by 50-80 percent.[25]

Fidelity of implementation appears to be extremely important for this program. Specifically, one of the randomized controlled trials of the program showed that when the home visits are carried out by paraprofessionals rather than nurses – holding all other details the same – the program is only marginally effective. Furthermore, a number of other home visitation programs for low-income families, designed for different purposes and using different protocols, have been shown in randomized controlled trials to be ineffective.[26]

B. When implementing an evidence-based intervention, it may be important to collect outcome data to check whether its effects in your schools differ greatly from what the evidence predicts.

Collecting outcome data is important because it is always possible that slight differences in implementation or setting between your schools or classrooms and those in the studies could lead to substantially different outcomes. So, for example, if you implement an evidence-based reading program in a particular group of schools or classrooms, you may wish to identify a comparison group of schools or classrooms, roughly matched in reading skills and demographic characteristics, that is not using the program. Tracking reading test scores for the two groups over time, while perhaps not fully meeting the guidelines for "possible" evidence described above, may still give you a sense of whether the program is having effects that are markedly different from what the evidence predicts.

BEST COPY AVAILABLE

Appendix A:
Where to find evidence-based interventions

The following web sites can be useful in finding evidence-based educational interventions. These sites use varying criteria for determining which interventions are supported by evidence, but all distinguish between randomized controlled trials and other types of supporting evidence. We recommend that, in navigating these web sites, you use this Guide to help you make independent judgments about whether the listed interventions are supported by "strong" evidence, "possible" evidence, or neither.

The What Works Clearinghouse (http://www.w-w-c.org/) established by the U.S. Department of Education's Institute of Education Sciences to provide educators, policymakers, and the public with a central, independent, and trusted source of scientific evidence of what works in education.

The Promising Practices Network (http://www.promisingpractices.net/) web site highlights programs and practices that credible research indicates are effective in improving outcomes for children, youth, and families.

Blueprints for Violence Prevention (http://www.colorado.edu/cspv/blueprints/index.html) is a national violence prevention initiative to identify programs that are effective in reducing adolescent violent crime, aggression, delinquency, and substance abuse.

The International Campbell Collaboration (http://www.campbellcollaboration.org/Fralibrary.html) offers a registry of systematic reviews of evidence on the effects of interventions in the social, behavioral, and educational arenas.

Social Programs That Work (http://www.excelgov.org/displayContent.asp?Keyword=prppcSocial) offers a series of papers developed by the Coalition for Evidence-Based Policy on social programs that are backed by rigorous evidence of effectiveness.

BEST COPY AVAILABLE

Appendix B:
Checklist to use in evaluating whether an intervention is backed by rigorous evidence

Step 1. **Is the intervention supported by "strong" evidence of effectiveness?**

A. **The quality of evidence needed to establish "strong" evidence: randomized controlled trials that are well-designed and implemented.** The following are key items to look for in assessing whether a trial is well-designed and implemented.

Key items to look for in the study's description of the intervention and the random assignment process

- ❑ **The study should clearly describe the intervention**, including: (i) who administered it, who received it, and what it cost; (ii) how the intervention differed from what the control group received; and (iii) the logic of how the intervention is supposed to affect outcomes (p. 5).
- ❑ **Be alert to any indication that the random assignment process may have been compromised.** (pp. 5-6).
- ❑ **The study should provide data showing that there are no systematic differences between the intervention and control groups prior to the intervention** (p. 6).

Key items to look for in the study's collection of outcome data

- ❑ **The study should use outcome measures that are "valid"** – – i.e., that accurately measure the true outcomes that the intervention is designed to affect (pp. 6-7).
- ❑ **The percent of study participants that the study has lost track of when collecting outcome data should be small, and should not differ between the intervention and control groups** (p. 7).
- ❑ **The study should collect and report outcome data even for those members of the intervention gr who do not participate in or complete the intervention** (p. 7).
- ❑ **The study should preferably obtain data on long-term outcomes of the intervention,** so that you can judge whether the intervention's effects were sustained over time (pp. 7-8).

Key items to look for in the study's reporting of results

- ❑ **If the study makes a claim that the intervention is effective, it should report (i) the size of the effect, and (ii) statistical tests showing the effect is unlikely to be the result of chance** (pp. 8-9).

BEST COPY AVAILABLE

- ❑ **A study's claim that the intervention's effect on a subgroup (e.g., Hispanic students) is different than its effect on the overall population in the study should be treated with caution** (p. 9).

- ❑ **The study should report the intervention's effects on all the outcomes that the study measured, not just those for which there is a positive effect.** (p. 9).

B. **Quantity of evidence needed to establish "strong" evidence of effectiveness** (p. 10).

- ❑ **The intervention should be demonstrated effective, through well-designed randomized controlled trials, in more than one site of implementation;**

- ❑ **These sites should be typical school or community settings**, such as public school classrooms taught by regular teachers; and

- ❑ **The trials should demonstrate the intervention's effectiveness in school settings similar to yours**, before you can be confident it will work in your schools/classrooms.

Step 2. **If the intervention is not supported by "strong" evidence, is it nevertheless supported by "possible" evidence of effectiveness?**

This is a judgment call that depends, for example, on the extent of the flaws in the randomized trials of the intervention and the quality of any nonrandomized studies that have been done. The following are a few factors to consider in making these judgments.

A. **Circumstances in which a comparison-group study can constitute "possible" evidence:**

- ❑ **The study's intervention and comparison groups should be very closely matched** in academic achievement levels, demographics, and other characteristics prior to the intervention (pp. 11-12).

- ❑ **The comparison group should not be comprised of individuals who had the option to participate in the intervention but declined** (p. 12).

- ❑ **The study should preferably choose the intervention/comparison groups and outcome measures "prospectively"** – i.e., *before* the intervention is administered (p. 12).

- ❑ **The study should meet the checklist items listed above for a well-designed randomized controlled trial** (other than the item concerning the random assignment process). That is, the study should use valid outcome measures, report tests for statistical significance, and so on (pp. 16-17).

B. **Studies that do *not* meet the threshold for "possible" evidence of effectiveness include:** (i) pre-post studies (p. 2); (ii) comparison-group studies in which the intervention and comparison groups are not well-matched; and (iii) "meta-analyses" that combine the results of individual studies which do not themselves meet the threshold for "possible" evidence (p. 13).

Step 3. **If the intervention is backed by neither "strong" nor "possible" evidence, one may conclude that it is not supported by meaningful evidence of effectiveness.**

BEST COPY AVAILABLE

References

[1] Evidence from randomized controlled trials, discussed in the following journal articles, suggests that one-on-one tutoring of at-risk readers by a well-trained tutor yields an effect size of about 0.7. This means that the average tutored student reads more proficiently than approximately 75 percent of the untutored students in the control group. Barbara A. Wasik and Robert E. Slavin, "Preventing Early Reading Failure With One-To-One Tutoring: A Review of Five Programs," *Reading Research Quarterly*, vol. 28, no. 2, April/May/June 1993, pp. 178-200 (the three programs evaluated in randomized controlled trials produced effect sizes falling mostly between 0.5 and 1.0). Barbara A. Wasik, "Volunteer Tutoring Programs in Reading: A Review," *Reading Research Quarterly*, vol. 33, no. 3, July/August/ September 1998, pp. 266-292 (the two programs using well-trained volunteer tutors that were evaluated in randomized controlled trials produced effect sizes of 0.5 to 1.0, and .50, respectively). Patricia F. Vadasy, Joseph R. Jenkins, and Kathleen Pool, "Effects of Tutoring in Phonological and Early Reading Skills on Students at Risk for Reading Disabilities, *Journal of Learning Disabilities*, vol. 33, no. 4, July/ August 2000, pages 579-590 (randomized controlled trial of a program using well-trained nonprofessional tutors showed effect sizes of 0.4 to 1.2).

[2] Gilbert J. Botvin et. al., "Long-Term Follow-up Results of a Randomized Drug Abuse Prevention Trial in a White, Middle-class Population," *Journal of the American Medical Association*, vol. 273, no. 14, April 12, 1995, pp. 1106-1112. Gilbert J. Botvin with Lori Wolfgang Kantor, "Preventing Alcohol and Tobacco Use Through Life Skills Training: Theory, Methods, and Empirical Findings," *Alcohol Research and Health*, vol. 24, no. 4, 2000, pp. 250-257.

[3] Frederick Mosteller, Richard J. Light, and Jason A. Sachs, "Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size," *Harvard Education Review*, vol. 66, no. 4, winter 1996, pp. 797-842. The small classes averaged 15 students; the regular-sized classes averaged 23 students.

[4] These are the findings specifically of the randomized controlled trials reviewed in "Teaching Children To Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction," *Report of the National Reading Panel*, 2000.

[5] Frances A. Campbell et. al., "Early Childhood Education: Young Adult Outcomes From the Abecedarian Project," *Applied Developmental Science*, vol. 6, no. 1, 2002, pp. 42-57. Craig T. Ramey, Frances A. Campbell, and Clancy Blair, "Enhancing the Life Course for High-Risk Children: Results from the Abecedarian Project," in *Social Programs That Work*, edited by Jonathan Crane (Russell Sage Foundation, 1998), pp. 163-183.

[6] For example, randomized controlled trials showed that (i) welfare reform programs that emphasized short-term job-search assistance and encouraged participants to find work quickly had larger effects on employment, earnings, and welfare dependence than programs that emphasized basic education; (ii) the work-focused programs were also much less costly to operate; and (iii) welfare-to-work programs often reduced net government expenditures. The trials also identified a few approaches that were particularly successful. See, for example, Manpower Demonstration Research Corporation, *National Evaluation of Welfare-to-Work Strategies: How Effective Are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs* (U.S. Department of Health and Human Services and U.S. Department of Education, November 2001). These valuable findings were a key to the political consensus behind the 1996 federal welfare reform legislation and its strong work requirements, according to leading policymakers — including Ron Haskins, who in 1996 was the staff director of the House Ways and Means Subcommittee with jurisdiction over the bill.

[7] See, for example, the Food and Drug Administration's standard for assessing the effectiveness of pharmaceutical drugs and medical devices, at 21 C.F.R. §314.126. See also, "The Urgent Need to Improve Health Care Quality," Consensus statement of the Institute of Medicine National Roundtable on Health Care Quality, *Journal of the American Medical Association*, vol. 280, no. 11, September 16, 1998, p. 1003; and Gary Burtless, "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, vol. 9, no. 2, spring 1995, pp. 63-84.

[8] Robert G. St. Pierre et. al., "Improving Family Literacy: Findings From the National Even Start Evaluation," Abt Associates, September 1996.

[9] Jean Baldwin Grossman, "Evaluating Social Policies: Principles and U.S. Experience," *The World Bank Research Observer*, vol. 9, no. 2, July 1994, pp. 159-181.

[10] Roberto Agodini and Mark Dynarski, "Are Experiments the Only Option? A Look at Dropout Prevention Programs," Mathematica Policy Research, Inc., August 2001, at http://www.mathematica-mpr.com/PDFs/redirect.asp?strSite=experonly.pdf.

[11] Elizabeth Ty Wilde and Rob Hollister, "How Close Is Close Enough? Testing Nonexperimental Estimates of Impact against Experimental Estimates of Impact with Education Test Scores as Outcomes," Institute for Research on Poverty Discussion paper, no. 1242-02, 2002, at http://www.ssc.wisc.edu/irp/.

BEST COPY AVAILABLE

[12]Howard S. Bloom et. al., "Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" MDRC Working Paper on Research Methodology, June 2002, at http://www.mdrc.org/ResearchMethodologyPprs.htm. James J. Heckman, Hidehiko Ichimura, and Petra E. Todd, "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, vol. 64, no. 4, 1997, pp. 605-654. Daniel Friedlander and Philip K. Robins, "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods," *American Economic Review*, vol. 85, no. 4, September 1995, pp. 923-937; Thomas Fraker and Rebecca Maynard, "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *Journal of Human Resources*, vol. 22, no. 2, spring 1987, pp. 194-227; Robert J. LaLonde, "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," *American Economic Review*, vol. 176, no. 4, September 1986, pp. 604-620.

[13] This literature, including the studies listed in the three preceding endnotes, is systematically reviewed in Steve Glazerman, Dan M. Levy, and David Myers, "Nonexperimental Replications of Social Experiments: A Systematic Review," Mathematica Policy Research discussion paper, no. 8813-300, September 2002. The portion of this review addressing labor market interventions is published in "Nonexperimental versus Experimental Estimates of Earnings Impact," *The American Annals of Political and Social Science*, vol. 589, September 2003.

[14] J.E. Manson et. al, "Estrogen Plus Progestin and the Risk of Coronary Heart Disease," *New England Journal of Medicine*, August 7, 2003, vol. 349, no. 6, pp. 519-522. *International Position Paper on Women's Health and Menopause: A Comprehensive Approach*, National Heart, Lung, and Blood Institute of the National Institutes of Health, and Giovanni Lorenzini Medical Science Foundation, NIH Publication No. 02-3284, July 2002, pp. 159-160. Stephen MacMahon and Rory Collins, "Reliable Assessment of the Effects of Treatment on Mortality and Major Morbidity, II: Observational Studies," *The Lancet*, vol. 357, February 10, 2001, p. 458. Sylvia Wassertheil-Smoller et. al., "Effect of Estrogen Plus Progestin on Stroke in Postmenopausal Women – The Women's Health Initiative: A Randomized Controlled Trial, *Journal of the American Medical Association*, May 28, 2003, vol. 289, no. 20, pp. 2673-2684.

[15] Howard S. Bloom, "Sample Design for an Evaluation of the Reading First Program," an MDRC paper prepared for the U.S. Department of Education, March 14, 2003. Robert E. Slavin, "Practical Research Designs for Randomized Evaluations of Large-Scale Educational Interventions: Seven Desiderata," paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 2003.

[16] The "standardized effect size" is calculated as the difference in the mean outcome between the treatment and control groups, divided by the pooled standard deviation.

[17] Rory Collins and Stephen MacMahon, "Reliable Assessment of the Effects of Treatment on Mortality and Major Morbidity, I: Clinical Trials," *The Lancet*, vol. 357, February 3, 2001, p. 375.

[18] Robinson G. Hollister, "The Growth of After-School Programs and Their Impact," paper commissioned by the Brookings Institution's Roundtable on Children, February 2003, at http://www.brook.edu/dybdocroot/views/papers/sawhill/20030225.pdf. Myles Maxfield, Allen Schirm, and Nuria Rodriguez-Planas, "The Quantum Opportunity Program Demonstration: Implementation and Short-Term Impacts," Mathematica Policy Research (no. 8279-093), August 2003.

[19] *Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drugs and Biological Products*, Food and Drug Administration, May 1998, pp. 2-5

[20] Robert J. Temple, Director of the Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration, quoted in Gary Taubes, "Epidemiology Faces Its Limits," *Science*, vol. 269, issue 5221, p. 169.

[21] Debra Viadero, "Researchers Debate Impact of Tests," *Education Week*, vol. 22, no. 21, February 5, 2003, page 1.

[22] E. Barrett-Connor and D. Grady, "Hormone Replacement Therapy, Heart Disease, and Other Considerations," *Annual Review of Public Health*, vol. 19, 1998, pp. 55-72.

[23] Frederick Mosteller, Richard J. Light, and Jason A. Sachs, op. cit., no. 3.

[24] Brian Stecher et. all, "Class-Size Reduction in California: A Story of Hope, Promise, and Unintended Consequences," *Phi Delta Kappan*, Vol. 82, Iss. 9, May 2001, pp. 670-674.

[25] David L. Olds et. al., "Long-term Effects of Nurse Home Visitation on Children's Criminal and Antisocial Behavior: 15-Year Follow-up of a Randomized Controlled Trial," *Journal of the American Medical Association*, vol. 280, no. 14, October 14, 1998, pp. 1238-1244. David L. Olds et. al., "Long-term Effects of Home Visitation on Maternal Life Course and Child Abuse and Neglect: 15-Year Follow-up of a Randomized Trial," *Journal of the American Medical Association*, vol. 278, no. 8, pp. 637-643. David L. Olds et. al, "Home Visiting By Paraprofessionals and By Nurses: A Randomized, Controlled Trial," *Pediatrics*, vol. 110, no. 3, September 2002, pp. 486-496. Harriet Kitzman et. al., "Effect of Prenatal and Infancy Home Visitation by Nurses on Pregnancy Outcomes, Childhood Injuries, and Repeated Childbearing," *Journal of the American Medical Association*, vol. 278, no. 8, August 27, 1997, pp. 644-652.

[26] For example, see Robert G. St. Pierre et. al., op. cit., no. 8; Karen McCurdy, "Can Home Visitation Enhance Maternal Social Support?" *American Journal of Community Psychology*, vol. 29, no. 1, 2001, pp. 97-112.

BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
X	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").